BGGN-213: FOUNDATIONS OF BIOINFORMATICS

The find-a-gene project assignment
http://thegrantlab.org/bggn213/
Dr. Barry Grant

Overview:

The find-a-gene project is a required assignment for BGGN-213. You should prepare a written report in **PDF** format that has responses to each question labeled **[Q1] - [Q10]** below. You may wish to consult the scoring rubric at the end of this document and the example report provided online.

The objective with this assignment is for you to demonstrate your grasp of database searching, sequence analysis, structure analysis and the R environment that we have covered in class.

Due Date:

Your responses to questions Q1-Q4 are due at the beginning of **Week 5**. Note that these answers can be obtained very quickly (at best within 10 or 15 minutes), so if you don't succeed at first, just keep trying.

The complete assignment, including responses to all questions, is due at the beginning of **Week 10**. Late responses will not be accepted under any circumstances.

Submission instructions:

Submit your PDF document to GradeScope as directed on our class website. Please do make sure your document is in PDF format and named something like `BGGN213_F20_[yourUCSDname].pdf` for example, my document would be named `BGGN213_F20_bjgrant.pdf`

Be sure to include your UCSD email and PID number on the first page of your report.

Submit your preliminary report with answers to Q1-Q4 at the beginning of **week 5** so we can determine if you have found a novel gene. Submit this preliminary report as one document with screen shots of the results inserted appropriately.

See the demonstration report linked to on the course website for an example of format. I will email you my decision; proceed with subsequent questions only after we are sure you have found a novel gene.

For the final report add your results for Q5-Q10 to the preliminary report and submit a final document containing the results for all questions. Please do not submit only Q5-Q10 answers as the final report.

Name: Jie Zhang
Email:jiz396@ucsd.edu
PID: A13814778

Questions:

[Q1] Tell me the name of a protein you are interested in. Include the species and the accession number. This can be a human protein or a protein from any other species as long as its function is known.

If you do not have a favorite protein, select human RBP4 or KIF11. Do not use beta globin as this is in the worked example report that I provide you with online.

Name: Sodium-dependent dopamine transporter
Species: Homo sapiens – human (Taxid:9606)
Accession: NP_001035.1
Function: A membrane dopamine transporter

[Q2] Perform a BLAST search against a DNA database, such as a database consisting of genomic DNA or ESTs. The BLAST server can be at NCBI or elsewhere. Include details of the BLAST method used, database searched and any limits applied (e.g. Organism).

Also include the output of that BLAST search in your document. If appropriate, change the font to `Courier size 10` so that the results are displayed neatly. You can also screen capture a BLAST output (e.g. alt print screen on a PC or on a MAC press ⌘-shift-4. The pointer becomes a bulls eye. Select the area you wish to capture and release. The image is saved as a file called `Screen Shot [].png` in your Desktop directory). It is **not** necessary to print out all of the blast results if there are many pages.

Method: TBLASTN search against nematode ESTs
Database: Expressed Sequence Tags (est)
Organism: non-homo sapiens

Job Title	NP_001035:sodium-dependent dopamine transporter...
RID	N5P4HJW8013 *Search expires on 10-23 01:36 am* Download All
Program	TBLASTN Citation
Database	est See details
Query ID	NP_001035.1
Description	sodium-dependent dopamine transporter [Homo sapiens]
Molecule type	amino acid
Query Length	620
Other reports	

Filter Results

Organism *only top 20 will appear* exclude

Type common name, binomial, taxid or group name

+ Add organism

Percent Identity to **E value** to **Query Coverage** to

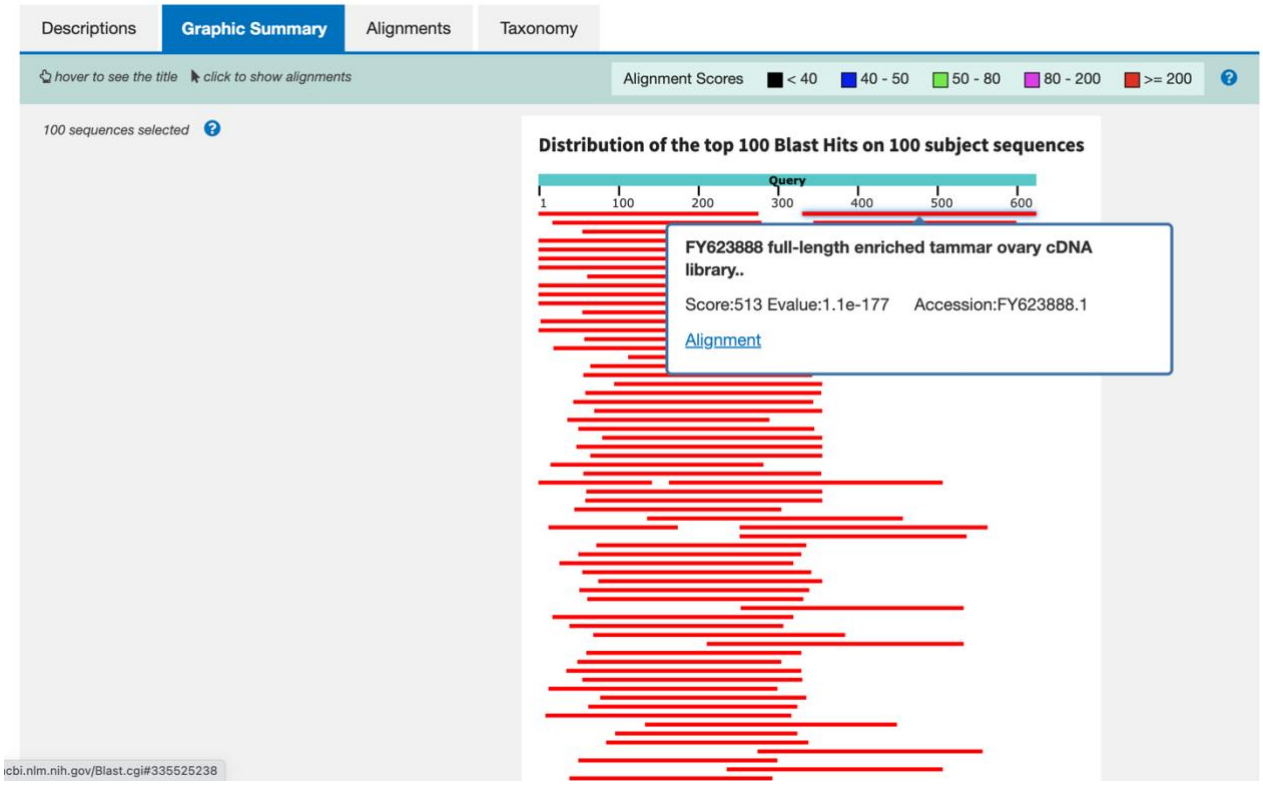
Filter Reset

Descriptions | Graphic Summary | Alignments | Taxonomy

Sequences producing significant alignments Download Select columns Show 100

select all *100 sequences selected* GenBank Graphics

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
FY623888 full-length enriched tammar ovary cDNA library Notamacropus eugenii cDNA clone MEOC-031D20 3',...	Notamacropus e...	513	513	46%	1e-177	84.88%	945	FY623888.1
AGENCOURT_10321515 NIH_MGC_144 Mus musculus cDNA clone IMAGE:6588897 5', mRNA sequence	Mus musculus	474	474	44%	5e-162	85.04%	1020	BU559951.1
IMMUNEF045808I17 POSSUM_01-C-POSSUM-IMMUNE-2KB Trichosurus vulpecula cDNA clone 10610210666...	Trichosurus vulp...	458	458	42%	3e-156	84.29%	900	DY612490.1
BL-76371 Nilaparvata lugens illumina library Nilaparvata lugens cDNA 5', mRNA sequence	Nilaparvata lugens	468	468	88%	6e-156	40.88%	1745	HS494205.1
BRAINSTEM4_F11.ab1 Korean native pig Brainstem by Oligo-capping method Sus scrofa cDNA 5', mRNA seque...	Sus scrofa	440	440	40%	6e-150	83.40%	764	GT889010.1
FQ140917 Rattus norvegicus 11-12 days foetus Sprague-Dawley Rattus norvegicus cDNA clone TL0ADA33YI04 ...	Rattus norvegicus	440	440	38%	1e-149	88.28%	816	FQ140917.1
nac24h04.y1 Dog eye lens. Unnormalized (nac) Canis lupus familiaris cDNA clone nac24h04 5', mRNA sequence	Canis lupus famil...	428	428	35%	1e-145	95.85%	667	DN866952.1
FQ144867 Rattus norvegicus 11-12 days foetus Sprague-Dawley Rattus norvegicus cDNA clone TL0ADA24YE17...	Rattus norvegicus	421	421	37%	9e-142	88.26%	848	FQ144867.1
FQ141251 Rattus norvegicus 11-12 days foetus Sprague-Dawley Rattus norvegicus cDNA clone TL0ADA32YK18...	Rattus norvegicus	414	414	36%	1e-139	88.89%	823	FQ141251.1
FQ143758 Rattus norvegicus 11-12 days foetus Sprague-Dawley Rattus norvegicus cDNA clone TL0ADA27YD06...	Rattus norvegicus	379	379	34%	4e-126	87.20%	732	FQ143758.1



>FY623888 full-length enriched tammar ovary cDNA library Notamacropus eugenii cDNA clone MEOC-031D20 3', mRNA sequence
Sequence ID: FY623888.1 Length: 945
Range 1: 73 to 945

```
Score:513 bits(1322), Expect:1e-177,
Method:Compositional matrix adjust.,
Identities:261/291(90%), Positives:277/291(95%), Gaps:0/291(0%)

Query  330  IAFSSYNKFTNNCYRDAIVTTSINsltsfssgfvvfsflgYMAQKHSVPIGDVAKDGPGL  389
            IAFSSYNKFTNNCYRDAI+TTS+NSLTSF SGFV+FSFLGYM+Q+H+VPIGDVAKDGPGL
Sbjct  945  IAFSSYNKFTNNCYRDAIITTSVNSLTSFFSGFVIFSFLGYMSQEHNVPIGDVAKDGPGL  766

Query  390  IFIIYPEAIATLPLSSAWAVVFFIMLLTLGIDSAMGGMESVITGLIDEFQLLHRHRELFT  449
            IFIIYPEAIATLPLSSAWAVVFFIMLLTLGIDSAMGGMESVITGLIDEF+ LHRHRELFT
Sbjct  765  IFIIYPEAIATLPLSSAWAVVFFIMLLTLGIDSAMGGMESVITGLIDEFKFLHRHRELFT  586

Query  450  LFIVLATFLLSLFCVTNGGIYVFTLLDHFAAGTSILFGVLIEAIGVAWFYGVGQFSDDIQ  509
            LFIVL+TFL SLFCVTNGGIYVFTLLDHFAAGTSILFGVLIEAIGVAWFYGVGQFSDDI+
Sbjct  585  LFIVLSTFLTSLFCVTNGGIYVFTLLDHFAAGTSILFGVLIEAIGVAWFYGVGQFSDDIK  406

Query  510  QMTGQRPSLYWRLCWKLVSPCFLLFVVVVSIVTFRPPHYGAYIFPDWANALGWVIATSSM  569
            QM G+RP LYWRLCWK VSPCFLLFVVVVSIVTFRPP+YG YIFPDWAN +GW+IATSSM
Sbjct  405  QMIGRRPGLYWRLCWKFVSPCFLLFVVVVSIVTFRPPNYGTYIFPDWANVVGWIIATSSM  226

Query  570  AMVPIYAAYKFCSLPGSFREKLAYAIAPEKDRELVDRGEVRQFTLRHWLKV  620
            AMVPIYA YKFCSLPGSFR+KLAYAI PEK+  LV+ GEVRQFTLRHWL V
Sbjct  225  AMVPIYATYKFCSLPGSFRKKLAYAITPEKEHGLVENGEVRQFTLRHWLMV  73
```

[Q3] Gather information about this "novel" **protein**. At a minimum, show me the protein sequence of the "novel" protein as displayed in your BLAST results from [Q2] as FASTA format (you can copy and paste the aligned sequence subject lines from your BLAST result page if necessary) or translate your novel DNA sequence using a tool called EMBOSS Transeq at the EBI. Don't forget to translate all six reading frames; the ORF (open reading frame) is likely to be the longest sequence without a stop codon. It may not start with a methionine if you don't have the complete coding region. Make sure the sequence you provide includes a header/subject line and is in traditional FASTA format.

```
>A Notamacropus eugenii protein

IAFSSYNKFTNNCYRDAIITTSVNSLTSFFSGFVIFSFLGYMSQEHNVPIGDVAKDGPGLIFIIYPEAIATLPLSSAWAV
VFFIMLLTLGIDSAMGGMESVITGLIDEFKFLHRHRELFTLFIVLSTFLTSLFCVTNGGIYVFTLLDHFAAGTSILFGVL
IEAIGVAWFYGVGQFSDDIKQMIGRRPGLYWRLCWKFVSPCFLLFVVVVSIVTFRPPNYGTYIFPDWANVVGWIIATSSM
MVPIYATYKFCSLPGSFRKKLAYAITPEKEHGLVENGEVRQFTLRHWLMV
```

Here, tell me the name of the novel protein, and the species from which it derives. It is very unlikely (but still definitely possible) that you will find a novel gene from an organism such as *S. cerevisiae*, human or mouse, because those genomes have already been thoroughly annotated. It is more likely that you will discover a new gene in a genome that is currently being sequenced, such as bacteria or plants or protozoa.

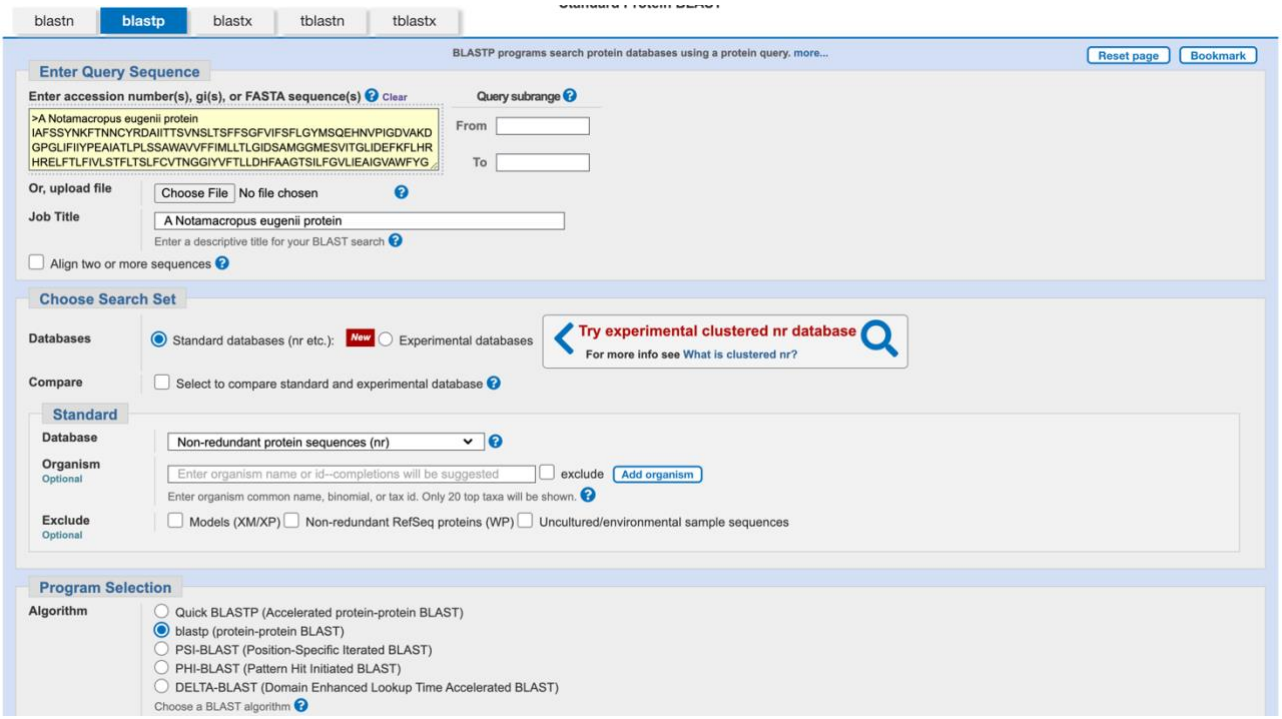
Name: sodium-dependent dopamine transporter

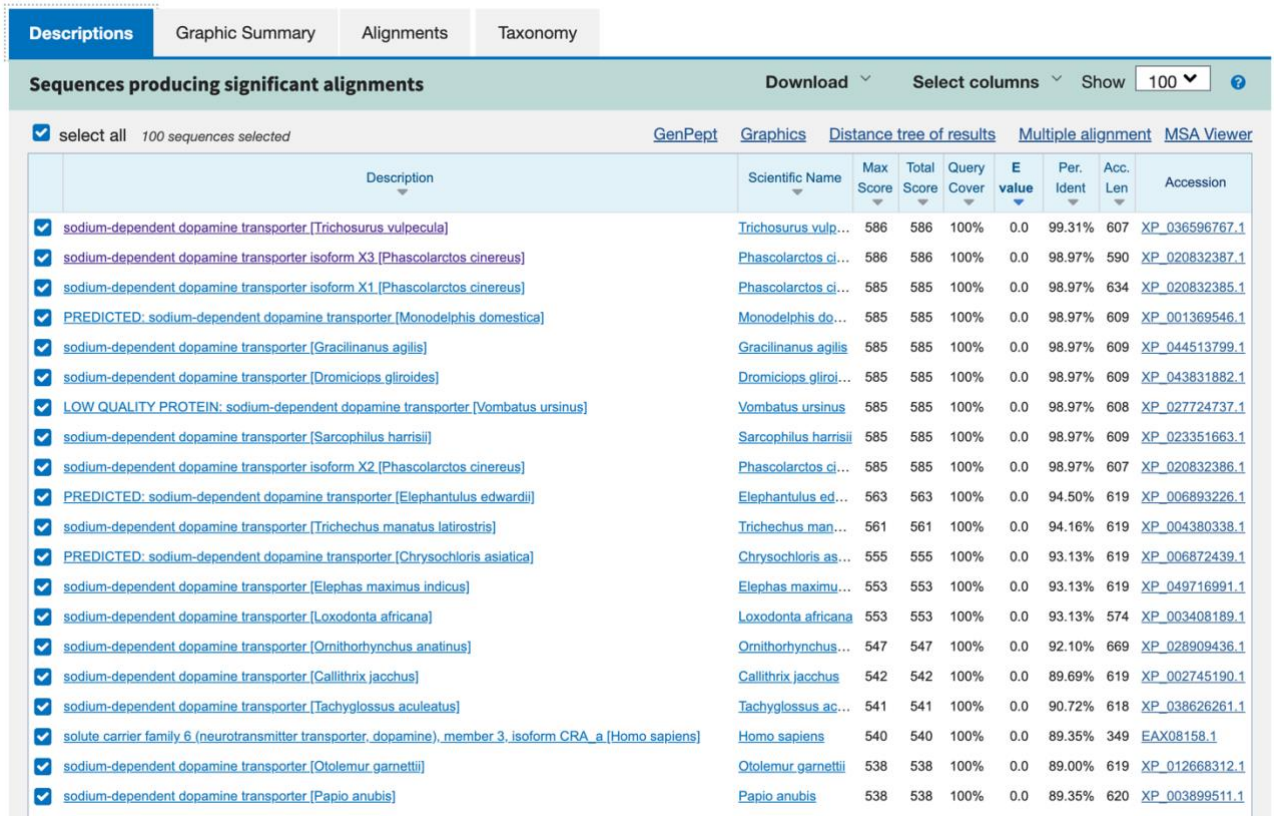
Species: Notamacropus eugenii (tammar wallaby)

Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Metatheria; Diprotodontia; Macropodidae; Notamacropus.

[Q4] Prove that this gene, and its corresponding protein, are novel. For the purposes of this project, "novel" is defined as follows. Take the protein sequence (your answer to [Q3]), and use it as a query in a blastp search of the nr database at NCBI.

- If there is a match with 100% amino acid identity to a protein in the database, from the same species, then your protein is NOT novel (even if the match is to a protein with a name such as "unknown"). Someone has already found and annotated this sequence, and assigned it an accession number.
- If the top match reported has less than 100% identity, then it is likely that your protein is novel, and you have succeeded.
- If there is a match with 100% identity, but to a different species than the one you started with, then you have likely succeeded in finding a novel gene.
- If there are no database matches to the original query from [Q1], this indicates that you have partially succeeded: yes, you may have found a new gene, but no, it is not actually homologous to the original query. You should probably start over.





Descriptions | Graphic Summary | Alignments | Taxonomy

Sequences producing significant alignments — Download | Select columns | Show 100

select all *100 sequences selected* — GenPept | Graphics | Distance tree of results | Multiple alignment | MSA Viewer

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
sodium-dependent dopamine transporter [Trichosurus vulpecula]	Trichosurus vulp...	586	586	100%	0.0	99.31%	607	XP_036596767.1
sodium-dependent dopamine transporter isoform X3 [Phascolarctos cinereus]	Phascolarctos ci...	586	586	100%	0.0	98.97%	590	XP_020832387.1
sodium-dependent dopamine transporter isoform X1 [Phascolarctos cinereus]	Phascolarctos ci...	585	585	100%	0.0	98.97%	634	XP_020832385.1
PREDICTED: sodium-dependent dopamine transporter [Monodelphis domestica]	Monodelphis do...	585	585	100%	0.0	98.97%	609	XP_001369546.1
sodium-dependent dopamine transporter [Gracilinanus agilis]	Gracilinanus agilis	585	585	100%	0.0	98.97%	609	XP_044513799.1
sodium-dependent dopamine transporter [Dromiciops gliroides]	Dromiciops gliroi...	585	585	100%	0.0	98.97%	609	XP_043831882.1
LOW QUALITY PROTEIN: sodium-dependent dopamine transporter [Vombatus ursinus]	Vombatus ursinus	585	585	100%	0.0	98.97%	608	XP_027724737.1
sodium-dependent dopamine transporter [Sarcophilus harrisii]	Sarcophilus harrisii	585	585	100%	0.0	98.97%	609	XP_023351663.1
sodium-dependent dopamine transporter isoform X2 [Phascolarctos cinereus]	Phascolarctos ci...	585	585	100%	0.0	98.97%	607	XP_020832386.1
PREDICTED: sodium-dependent dopamine transporter [Elephantulus edwardii]	Elephantulus ed...	563	563	100%	0.0	94.50%	619	XP_006893226.1
sodium-dependent dopamine transporter [Trichechus manatus latirostris]	Trichechus man...	561	561	100%	0.0	94.16%	619	XP_004380338.1
PREDICTED: sodium-dependent dopamine transporter [Chrysochloris asiatica]	Chrysochloris as...	555	555	100%	0.0	93.13%	619	XP_006872439.1
sodium-dependent dopamine transporter [Elephas maximus indicus]	Elephas maximu...	553	553	100%	0.0	93.13%	619	XP_049716991.1
sodium-dependent dopamine transporter [Loxodonta africana]	Loxodonta africana	553	553	100%	0.0	93.13%	574	XP_003408189.1
sodium-dependent dopamine transporter [Ornithorhynchus anatinus]	Ornithorhynchus...	547	547	100%	0.0	92.10%	669	XP_028909436.1
sodium-dependent dopamine transporter [Callithrix jacchus]	Callithrix jacchus	542	542	100%	0.0	89.69%	619	XP_002745190.1
sodium-dependent dopamine transporter [Tachyglossus aculeatus]	Tachyglossus ac...	541	541	100%	0.0	90.72%	618	XP_038626261.1
solute carrier family 6 (neurotransmitter transporter, dopamine), member 3, isoform CRA_a [Homo sapiens]	Homo sapiens	540	540	100%	0.0	89.35%	349	EAX08158.1
sodium-dependent dopamine transporter [Otolemur garnettii]	Otolemur garnettii	538	538	100%	0.0	89.00%	619	XP_012668312.1
sodium-dependent dopamine transporter [Papio anubis]	Papio anubis	538	538	100%	0.0	89.35%	620	XP_003899511.1

The top results are different species, and the identity is not 100%

[Q5] Generate a multiple sequence alignment with your novel protein, your original query protein, and a group of other members of this family from different species. A typical number of proteins to use in a multiple sequence alignment for this assignment purpose is a minimum of 5 and a maximum of 20 - although the exact number is up to you. Include the multiple sequence alignment in your report. Use Courier font with a size appropriate to fit page width.

Side-note: Indicate your sequence in the alignment by choosing an appropriate name for each sequence in the input unaligned sequence file (i.e. edit the sequence file so that the species, or short common, names (rather than accession numbers) display in the output alignment and in the subsequent answers below). The goal in this step is to create an interesting an alignment for building a phylogenetic tree that illustrates species divergence.

```
>Notamacropus_eugenii protein(sequence taken from BLAST result)
IAFSSYNKFTNNCYRDAIITTSVNSLTSFFSGFVIFSFLGYMSQEHNVPIGDVAKDGPGLIFIIYPEAIATLPLSSAWAV
VFFIMLLTLGIDSAMGGMESVITGLIDEFKFLHRHRELFTLFIVLSTFLTSLFCVTNGGIYVFTLLDHFAAGTSILFGVL
IEAIGVAWFYGVGQFSDDIKQMIGRRPGLYWRLCWKFVSPCFLLFVVVVSIVTFRPPNYGTYIFPDWANVVGWIIATSSM
MVPIYATYKFCSLPGSFRKKLAYAITPEKEHGLVENGEVRQFTLRHWLMV

>Homo_sapien AAC50179.2|dopamine transporter
MSKSKCSVGLMSSVVAPAKEPNAVGPKEVELILVKEQNGVQLTSSTLTNPRQSPVEAQDRETWGKKIDFL
LSVIGFAVDLANVWRFPYLCYKNGGGAFLVPYLLFMVIAGMPLFYMELALGQFNREGAAGVWKICPILKG
VGFTVILISLYVGFFYNVIIAWALHYLFSSFTTELPWIHCNNSWNSPNCSDAHPGDSSGDSSGLNDTFGT
TPAAEYFERGVLHLHQSHGIDDLGPPRWQLTACLVLVIVLLYFSLWKGVKTSGKVVWITATMPYVVLTAL
LLRGVTLPGAIDGIRAYLSVDFYRLCEASVWIDAATQVCFSLGVGFGVLIAFSSYNKFTNNCYRDAIVTT
SINSLTSFSSGFVVFSFLGYMAQKHSVPIGDVAKDGPGLIFIIYPEAIATLPLSSAWAVVFFIMLLTLGI
DSAMGGMESVITGLIDEFQLLHRHRELFTLFIVLATFLLSLFCVTNGGIYVFTLLDHFAAGTSILFGVLI
EAIGVAWFYGVGQFSDDIQQMTGQRPSLYWRLCWKLVSPCFLLFVVVVSIVTFRPPHYGAYIFPDWANAL
```

GWVIATXSMAMVPIYAAYKFCSLPGSFREKLAYAIAPEKDRELVDRGEVRQFTLRHWLKV

>Bombyx_mori NP_001037362.2|sodium-dependent dopamine transporter
MALKTPTPGVVGERETWGKKVDFLLSVIGFAVDLANVWRFPYLCYKNGGGAFLVPYCIMLVVGGIPLFYM
ELALGQFHRKGAITCWGRLVPLFKGIGYAVVLIAFYVDFYYNVIIAWALRFFFASFTTMLPWTNCDNEWN
TPACRPFEAIWDVNRTRIRNTTSASLGIAPTTPYTSAASEYFNRAILELQGSEGLHDLGSVKWDMALCLL
AVYVICYFSLWKGISTSGKVVWFTALFPYAVLLILLVRGITLPGSATGIQYYLSPNFEAITQPQVWVDAA
TQVFFSLGPGFGVLLAYASYNKYHNNVYKDAILTSVINSATSFVAGFVIFSVLGYMAHASGRDVQDVATE
GPGLVFVVYPAAIATMPGSTFWALIFFMMLLTLGLDSSFGGSEAIITALSDEFPPIGRHRELFVACLFTL
YFFVGLASCTKGGFYFFQLLDRYAAGYSILIAVFFEAIAVSWIYGTERFCEDIRDMIGFRPGLYWRVCWR
FAAPSFLLFITAYGLLDYEPLQYENYIYPGWANALGWAIAGSSVMCIPTVAIYKLITTKGSFLERLRVLT
TPYADSERNGTVHNGMIVSESGGVRLTSAVQTPTTPQQPIGANIPAASAPTLASSPALV

>Drosophila NP_001261026.1| dopamine transporter, isoform B
MSPTGHISKSKTPTPHDNDNNSISDERETWSGKVDFLLSVIGFAVDLANVWRFPYLCYKNGGGAFLVPYG
IMLVVGGIPLFYMELALGQHNRKGAITCWGRLVPLFKGIGYAVVLIAFYVDFYYNVIIAWSLRFFFASFT
NSLPWTSCNNIWNTPNCRPFESQNASRVPVIGNYSDLYAMGNQSLLYNETYMNGSSLDTSAVGHVEGFQS
AASEYFNRYILELNRSEGIHDLGAIKWDMALCLLIVYLICYFSLWKGISTSGKVVWFTALFPYAVLLILL
IRGLTLPGSFLGIQYYLTPNFSAIYKAEVWVDAATQVFFSLGPGFGVLLAYASYNKYHNNVYKDALLTSF
INSATSFIAGFVIFSVLGYMAHTLGVRIEDVATEGPGLVFVVYPAAIATMPASTFWALIFFMMLLTLGLD
SSFGGSEAIITALSDEFPKIKRNRELFVAGLFSLYFVVGLASCTQGGFYFFHLLDRYAAGYSILVAVFFE
AIAVSWIYGTNRFSEDIRDMIGFPPGRYWQVCWRFVAPIFLLFITVYGLIGYEPLTYADYVYPSWANALG
WCIAGSSVVMIPAVAIFKLLSTPGSLRQRFTILTTPWRDQQSMAMVLNGVTTEVTVVRLTDTETAKEPVD
VXVRPVARFQIYYV

>Vibrio_cholerae TYC39863.1|Sodium-dependent dopamine transporter
MAQSNSRETFSSRLGFILAAAGAAVGLGNIWGFPTQAASNGGGAFLLVYLILIFVVAFPMLVVEMAIGRY
GQANPVDSMRSLTSQPAAKKVGGFVGWVGLSVPSAVLAFYSIVGGWIICFLLGAMTDLFGFTAASAWLKG
FSVERNLFGTLIFYVLTILIVQGGVKQGIERWSTRLMPALFVLFAVLFIYIMTQQGAWEGLKHYLIPDFE
KVWDRKLILAAMGQGFFSLTIGGCSMLIYGSYLSKKENLPKMAMSVTLVDTAVAFIAGLVVLPAMFVAMN
KGVQIYAQDGSLLSSDTLVFTVLPLMFDSLGLLGQIFAMVFFLLLTIAALTSSISMLECPVALVGERFNT
RRTPTSWVLGGLIALFSVVIVYNFGALFGLVATLATQYLQPTAALMFCLFGGWVWQRDAKMKELQAGFPE
LQQSLFGKIWPWYVKFVCPVLVATVIWASFG

>Pan_paniscus XP_003805237.1|sodium-dependent dopamine transporter isoform X2
MSKSKCSVGLMSSVVAPAKEPNAVGPKEVELILVKEQNGVQLTSSTLTNPRQSPVEAQDRETWGKKIDFL
LSVIGFAVDLANVWRFPYLCYKNGGGAFLVPYLLFMVIAGMPLFYMELALGQFNREGAAGVWKICPILKG
VGFTVILISLYVGFFYNVIIAWALHYLFSSFTTELPWIHCNNSWNSPNCSDAHPGDSSGDSSGLNDTFGT
TPAAEYFERGVLHLHQSHGIDDLGPPRWQLTACLVLVIVLLYFSLWKGVKTSGKVVWITATMPYVVLTAL
LLRGVTLPGAIDGIRAYLSVDFYRLCEASVWIDAATQVCFSLGVGFGVLIAFSSYNKFTNNCYRDAIVTT
SINSLTSFSSGFVVFSFLGYMAQKHSVPIGDVAKDGPGLIFIIYPEAIATLPLSSAWAVVFFIMLLTLGI
DSAMGGMESVITGLIDEFQLLHRHRELFTLFIVLATFLLSLFCVTNGGIYVFTLLDHFAAGTSILFGVLI
EAIGVAWFYGVGQFSDDIQQMTGQRPSLYWRLCWKLVSPCFLLFVVVVSIVTFRPPHYGAYIFPDWANAL
GWVIATSSMAMVPIYAAYKFCSLRGSFREKLAYAIAPEKDRELVDRGEVRQFTLRHWLKV

Vibrio_cholerae **MAQSN----------SR**
Notamacropus_eugenii **----------**
Homo_sapien **MSKSKCSVGLMSSVVAPAKEPNAVGPKEVELILVKEQNGVQLTSSTLTNPRQSPVEAQDR**
Pan_paniscus **MSKSKCSVGLMSSVVAPAKEPNAVGPKEVELILVKEQNGVQLTSSTLTNPRQSPVEAQDR**
Bombyx_mori **-------MALKTPTP------GVVG----------ER**
Drosophila **MSPTGHISKSKTPTPHDNDNNSISD----------ER**

Vibrio_cholerae **ETFSSRLGFILAAAGAAVGLGNIWGFPTQAASNGGGAFLLVYLILIFVVAFPMLVVEMAI**
Notamacropus_eugenii **----------**
Homo_sapien **ETWGKKIDFLLSVIGFAVDLANVWRFPYLCYKNGGGAFLVPYLLFMVIAGMPLFYMELAL**
Pan_paniscus **ETWGKKIDFLLSVIGFAVDLANVWRFPYLCYKNGGGAFLVPYLLFMVIAGMPLFYMELAL**
Bombyx_mori **ETWGKKVDFLLSVIGFAVDLANVWRFPYLCYKNGGGAFLVPYCIMLVVGGIPLFYMELAL**
Drosophila **ETWSGKVDFLLSVIGFAVDLANVWRFPYLCYKNGGGAFLVPYGIMLVVGGIPLFYMELAL**

Vibrio_cholerae **GRYGQANPVDSMRSLTSQPAAKKVGGFVGWVGLSVPSAVLAFYSIVGGWIICFLLGAMTD**

```
Notamacropus_eugenii      ------------------------------------------------------------
Homo_sapien               GQFNREGAAGVW-KIC--PILKGVGFTVILISLYVGF----FYNVIIAWALHYLFSSFTT
Pan_paniscus              GQFNREGAAGVW-KIC--PILKGVGFTVILISLYVGF----FYNVIIAWALHYLFSSFTT
Bombyx_mori               GQFHRKGAITCWGRLV--PLFKGIGYAVVLIAFYVDF----YYNVIIAWALRFFFASFTT
Drosophila                GQHNRKGAITCWGRLV--PLFKGIGYAVVLIAFYVDF----YYNVIIAWSLRFFFASFTN

Vibrio_cholerae           LFGFTA-----------------------------------------ASAWLKGFSVERNL
Notamacropus_eugenii      ------------------------------------------------------------
Homo_sapien               ELPWIHCNNSWNSPNC------------------------------SDAHPGDSSGDSSG
Pan_paniscus              ELPWIHCNNSWNSPNC------------------------------SDAHPGDSSGDSSG
Bombyx_mori               MLPWTNCDNEWNTPACRPFEA--------------IWDV-------NRTRIRNTTSASLG
Drosophila                SLPWTSCNNIWNTPNCRPFESQNASRVPVIGNYSDLYAMGNQSLLYNETYMNGSSLDTSA

Vibrio_cholerae           FGT--------------------------------------------LIFYVLTILIVQGGVKQ
Notamacropus_eugenii      ------------------------------------------------------------
Homo_sapien               LNDTFG-TTPAAEYFERGVLHLHQSHGIDDLGPPRWQLTACLVLVIVLLYFSLWKGVKTS
Pan_paniscus              LNDTFG-TTPAAEYFERGVLHLHQSHGIDDLGPPRWQLTACLVLVIVLLYFSLWKGVKTS
Bombyx_mori               IAPTTPYTSAASEYFNRAILELQGSEGLHDLGSVKWDMALCLLAVYVICYFSLWKGISTS
Drosophila                VGHVEGFQSAASEYFNRYILELNRSEGIHDLGAIKWDMALCLLIVYLICYFSLWKGISTS

Vibrio_cholerae           GIERWSTRLMPALFVLFAVLFIYIMTQQGAWEGLKHYLIPDFEKVWDRKLILAAMGQGFF
Notamacropus_eugenii      ------------------------------------------------------------
Homo_sapien               GKVVWITATMP--YVVLTALLLRGVTLPGAIDGIRAYLSVDFYRLCEASVWIDAATQVCF
Pan_paniscus              GKVVWITATMP--YVVLTALLLRGVTLPGAIDGIRAYLSVDFYRLCEASVWIDAATQVCF
Bombyx_mori               GKVVWFTALFP--YAVLLILLVRGITLPGSATGIQYYLSPNFEAITQPQVWVDAATQVFF
Drosophila                GKVVWFTALFP--YAVLLILLIRGLTLPGSFLGIQYYLTPNFSAIYKAEVWVDAATQVFF

Vibrio_cholerae           SLTIGGCSMLIYGSYLSKKENLPKMAMSVTLVDTAVAFIAGLVVLPAM-FVAMNKGVQI-
Notamacropus_eugenii      ---------IAFSSYNKFTNNCYRDAIITTSVNSLTSFFSGFVIFSFLGYMSQEHNVPIG
Homo_sapien               SLGVGFGVLIAFSSYNKFTNNCYRDAIVTTSINSLTSFSSGFVVFSFLGYMAQKHSVPIG
Pan_paniscus              SLGVGFGVLIAFSSYNKFTNNCYRDAIVTTSINSLTSFSSGFVVFSFLGYMAQKHSVPIG
Bombyx_mori               SLGPGFGVLLAYASYNKYHNNVYKDAILTSVINSATSFVAGFVIFSVLGYMAHASGRDVQ
Drosophila                SLGPGFGVLLAYASYNKYHNNVYKDALLTSFINSATSFIAGFVIFSVLGYMAHTLGVRIE
                                   : :.** .  :*  . *: .: ::: .:* :*:*::. : :::   .  :

Vibrio_cholerae           -YAQDGSLLSSDTLVFTVLPLMFDSLGLLGQIFAMVFFLLLTIAALTSSISMLECPVALV
Notamacropus_eugenii      DVAKDGP-----GLIFIIYPEAIATLP-LSSAWAVVFFIMLLTLGIDSAMGGMESVITGL
Homo_sapien               DVAKDGP-----GLIFIIYPEAIATLP-LSSAWAVVFFIMLLTLGIDSAMGGMESVITGL
Pan_paniscus              DVAKDGP-----GLIFIIYPEAIATLP-LSSAWAVVFFIMLLTLGIDSAMGGMESVITGL
Bombyx_mori               DVATEGP-----GLVFVVYPAAIATMP-GSTFWALIFFMMLLTLGLDSSFGGSEAIITAL
Drosophila                DVATEGP-----GLVFVVYPAAIATMP-ASTFWALIFFMMLLTLGLDSSFGGSEAIITAL
                            * :*.      *:* : *  : ::   .  :*::**::*   .: *::.  *. :: :

Vibrio_cholerae           GERFNTRRTPTSWVLGGLIALFSVVIVYNF--GALFGLVATLATQYLQPTAALMFCLFGG
Notamacropus_eugenii      IDEFKFLHRHRELFTLFIVLSTFLTSLFCVTNGGIY--VFTLLDHFAAGTSILFGVLIEA
Homo_sapien               IDEFQLLHRHRELFTLFIVLATFLLSLFCVTNGGIY--VFTLLDHFAAGTSILFGVLIEA
Pan_paniscus              IDEFQLLHRHRELFTLFIVLATFLLSLFCVTNGGIY--VFTLLDHFAAGTSILFGVLIEA
Bombyx_mori               SDEFPPIGRHRELFVACLFTLYFFVGLASCTKGGFY--FFQLLDRYAAGYSILIAVFFEA
Drosophila                SDEFPKIKRNRELFVAGLFSLYFVVGLASCTQGGFY--FFHLLDRYAAGYSILVAVFFEA
                           : *       . .   :.    .  :     *.::  .  *  .:    : *.  :: .

Vibrio_cholerae           ----WVWQRDAKMKELQ--AGFPELQQSLFGKIWPWYVKFVCPVLVATVI----------
Notamacropus_eugenii      IGVAWFYGVGQFSDDIKQMIGRRP------GLYWRLCWKFVSPCFLLFVVVVSIVTFRPP
Homo_sapien               IGVAWFYGVGQFSDDIQQMTGQRP------SLYWRLCWKLVSPCFLLFVVVVSIVTFRPP
Pan_paniscus              IGVAWFYGVGQFSDDIQQMTGQRP------SLYWRLCWKLVSPCFLLFVVVVSIVTFRPP
Bombyx_mori               IAVSWIYGTERFCEDIRDMIGFRP------GLYWRVCWRFAAPSFLLFITAYGLLDYEPL
Drosophila                IAVSWIYGTNRFSEDIRDMIGFPP------GRYWQVCWRFVAPIFLLFITVYGLIGYEPL
                              *.:      .::.   *         .  *    .:..* ::  :

Vibrio_cholerae           ---------WASF-----------------------------------------------
Notamacropus_eugenii      NYGTYIFPDWANVVGWIIATSSM-MVPIYATYKFCSLPGSFRKKLAYAITPEKEH---GL
Homo_sapien               HYGAYIFPDWANALGWVIATXSMAMVPIYAAYKFCSLPGSFREKLAYAIAPEKDR---EL
Pan_paniscus              HYGAYIFPDWANALGWVIATSSMAMVPIYAAYKFCSLRGSFREKLAYAIAPEKDR---EL
Bombyx_mori               QYENYIYPGWANALGWAIAGSSVMCIPTVAIYKLITTKGSFLERLRVLTTPYADSERNGT
```

```
Drosophila                TYADYVYPSWANALGWCIAGSSVVMIPAVAIFKLLSTPGSLRQRFTILTTPWRDQQSMAM
                                  **.

Vibrio_cholerae           ---G-------------------------------------------
Notamacropus_eugenii      VENG-----------------EVRQFTLRHWLM------------V
Homo_sapien               VDRG-----------------EVRQFTLRHWLK------------V
Pan_paniscus              VDRG-----------------EVRQFTLRHWLK------------V
Bombyx_mori               VHNGMIVSESGGVRLTSAVQTPTTPQQPIGANIPAASAPTLASSPALV
Drosophila                VLNG-VTTEVTVVRLTDT----ETAKEPVDVXVR----PVARFQIYYV
                             *
```

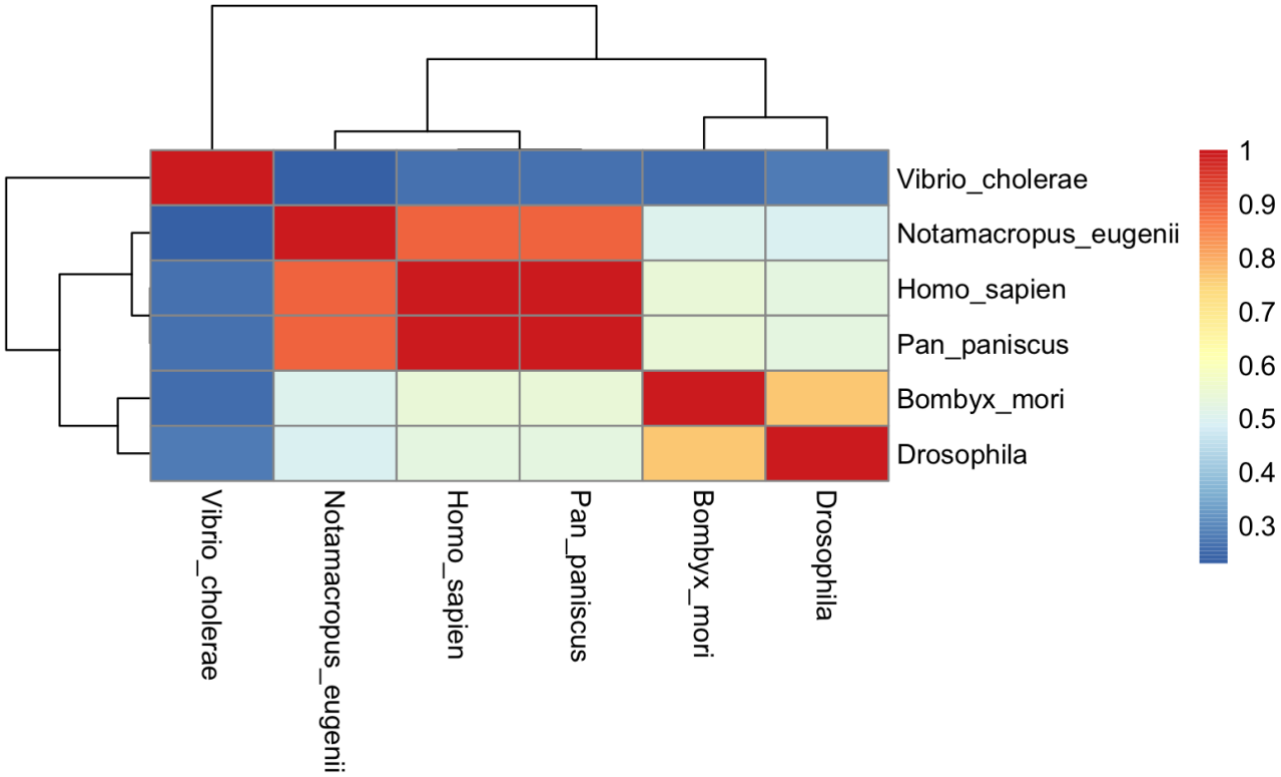
[Q6] Create a phylogenetic tree, using either a parsimony or distance-based approach. Bootstrapping and tree rooting are optional. Use "simple phylogeny" online from the EBI or any respected phylogeny program (such as MEGA, PAUP, or Phylip). Paste an image of your Cladogram or tree output in your report.

Using simply phylogeny from EBI and select Cladogram in form

[Q7] Generate a sequence identity based **heatmap** of your aligned sequences using R.

If necessary, convert your sequence alignment to the ubiquitous FASTA format (Seaview can read in clustal format and "Save as" FASTA format for example). Read this FASTA format alignment into R with the help of functions in the **Bio3D package**. Calculate a sequence identity matrix (again using a function within the Bio3D package). Then generate a heatmap plot and add to your report. Do make sure your labels are visible and not cut at the figure margins.

Heatmap is generated with pheatmap package



[Q8] Using R/Bio3D (or an online blast server if you prefer), search the main protein structure database for the most similar atomic resolution structures to your aligned sequences.

List the top 3 *unique* hits (i.e. not hits representing different chains from the same structure) along with their E-value and sequence identity to your query. Please also add annotation details of these

structures. For example include the annotation terms PDB identifier (structureId), Method used to solve the structure (experimental Technique), resolution (resolution), and source organism (source).

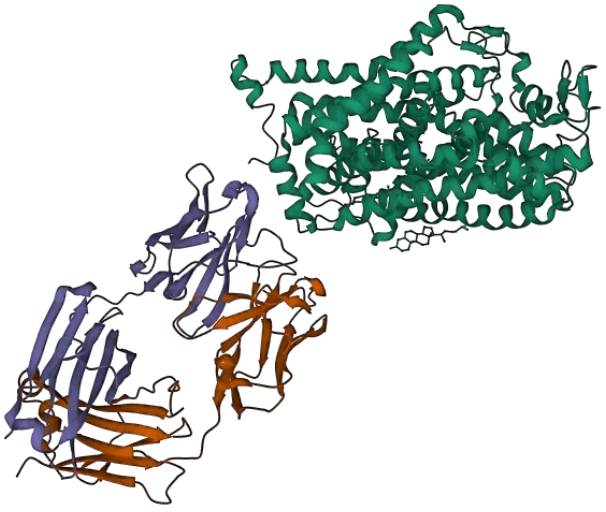
HINT: You can use a single sequence from your alignment or generate a consensus sequence from your alignment using the Bio3D function consensus(). The Bio3D functions blast.pdb(), plot.blast() and pdb.annotate() are likely to be of most relevance for completing this task. Note that the results of blast.pdb() contain the hits PDB identifier (or pdb.id) as well as E-value and identity. The results of pdb.annotate() contain the other annotation terms noted above.

Note that if your consensus sequence has lots of gap positions then it will be better to use an original sequence from the alignment for your search of the PDB. In this case you could chose the sequence with the highest identity to all others in your alignment by calculating the row-wise maximum from your sequence identity matrix.

ID	Technique	Resolution	Source	E-value	Identity
4XPG_A	**X-ray Diffraction**	**3.21 Å**	**Drosophila melanogaster**	**2e-80**	**51.27**
6VRH_A	**Electron Microscopy (Cyro-EM)**	**3.3 Å**	**Homo sapiens**	**5e-69**	**45.58**
7SK2_A	**Electron Microscopy (Cyro-EM)**	**3.82 Å**	**Homo sapiens**	**5e-67**	**45.79**

[Q9] Generate a molecular figure of one of your identified PDB structures using **VMD**. You can optionally highlight conserved residues that are likely to be functional. Please use a white or transparent background for your figure (i.e. not the default black).

The molecule is generated by Mol* Viewer. My computer doesn't support VMD. Th following structure is the 4XPG protein.



Based on sequence similarity. How likely is this structure to be similar to your "novel" protein?

The sequence similarity is around 50% so it is hard to conclude that the structure is very similar to the Notamacropus eugenii protein but at least over half is similar. I suspect the similarity mostly lies on dopamine transporter chain A which in the figure is colored green and corresponds to the Notamacropus eugenii transporter subject of this report. Purple and orange chains are ligand fragment heavy and light chain which might not exists in the Notamacropus eugenii protein.

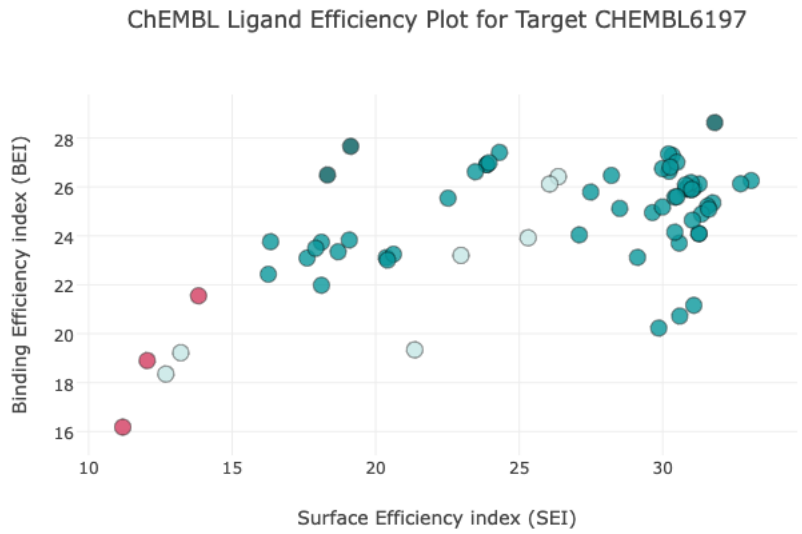
[Q10] Perform a " (https://www.ebi.ac.uk/chembl/) with your novel sequence. Are there any **Target Associated Assays** and **ligand efficiency data** reported that may be useful starting points for exploring potential inhibition of your novel protein?

Target report card:
https://www.ebi.ac.uk/chembl/target_report_card/CHEMBL6197/

Target Associated Assays: 1 Binding Assay (CHEMBL914034)
Binding assay description: Displacement of [3H]WIN-35428 from DAT in rhesus monkey caudate-putamen
thiabicyclo[3.2.1]octanes which are DAT inhibitors in search for medications for cocaine abuse.
https://www.ebi.ac.uk/chembl/assay_report_card/CHEMBL914034/

Duy-Phong Pham-Huu, Jeffrey R. Deschamps, Shanghao Liu, Bertha K. Madras, Peter C. Meltzer, Synthesis of 8-thiabicyclo[3.2.1]octanes and their binding affinity for the dopamine and serotonin transporters, Bioorganic & Medicinal Chemistry,
Volume 15, Issue 2, 2007, Pages 1067-1082, ISSN 0968-0896,
https://doi.org/10.1016/j.bmc.2006.10.016.

Ligand efficiency data:



Scoring Rubric:

[45 total points available]

Q1 (4 points)	
Protein name	1
Species	1
Accession number	1
Function known	1
Q2 (6 points)	
Blast method	1
Database searched	1
Limits applied	1
Search output list (top hits)	1
Alignment of choice	1
Evalue and other alignment stats	1
Q3 (3 points)	
Protein sequence of choice matches Subject above	1
Name in header	1
Species	1
Q4 (3 point)	
Blastp output list with identities & Evalue	1
Top alignment shown with alignment statistics	1

Results indicates a “novel” gene found	1

Q5 (3 points)

MSA labeled with useful names	1
MSA trimmed appropriately (i.e. no gap overhangs)	1
Pasted MSA fits report page width (i.e. font, format)	1

Q6 (1 point)

Figure illustrates sequence clustering pattern	1

Q7 (10 points)

Heatmap figure included in report	5
Heatmap is legible (i.e. no labels obscured)	5

Q8 (10 points)

PDB identifiers from multiple species reported	5
Annotation of PDB source, resolution and technique	4
Annotation of Evalue and Sequence Identity	1

Q9 (4 points)

Structure figure provided	2
Uses white background for molecular figure	1
Figure of high resolution (i.e. not just snapshot)	1

Q10 (1 point)

Evidence of ChEMBEL searches	1